Using Symbolic Knowledge in the UMLS to Disambiguate Words in Small Datasets with a Naïve Bayes Classifier.

Gondy Leroy[a*], Thomas C. Rindflesch[b]

[a]*School of Information Science, Claremont Graduate University, Claremont, CA, USA*
[b]*National Library of Medicine, Bethesda, MD, USA*

Abstract

Current approaches to word sense disambiguation use and combine various machine-learning techniques. Most refer to characteristics of the ambiguous word and surrounding words and are based on hundreds of examples. Unfortunately, developing large training sets is time-consuming. We investigate the use of symbolic knowledge to augment machine-learning techniques for small datasets. UMLS semantic types assigned to concepts found in the sentence and relationships between these semantic types form the knowledge base. A naïve Bayes classifier was trained for 15 words with 100 examples for each. The most frequent sense of a word served as the baseline. The effect of increasingly accurate symbolic knowledge was evaluated in eight experimental conditions. Performance was measured by accuracy based on 10-fold cross-validation. The best condition used only the semantic types of the words in the sentence. Accuracy was then on average 10% higher than the baseline; however, it varied from 8% deterioration to 29% improvement. In a follow-up evaluation, we noted a trend that the best disambiguation was found for words that were the least troublesome to the human evaluators.

Keywords: Artificial intelligence, machine learning, naïve Bayes, word sense disambiguation, Unified Medical Language System, UMLS, small datasets, symbolic knowledge

Introduction

Although many words we use in conversation and writing are ambiguous, we usually do not experience problems with interpreting these words in their context. This is, however, not easily accomplished with automated methods. Since this is an important issue for machine translation, information retrieval, thematic analysis, or any type of speech and text processing, many researchers have devoted time to word sense disambiguation (WSD). WSD techniques choose the correct sense for a word from a predefined set of available senses. Most existing techniques use the surrounding words and specific features of these to learn the correct sense of the ambiguous word. They are usually supervised and based on annotated dataset where the correct sense is indicated for each instance. Ide [1] provides an overview of WSD from the early years (1950's) through the late 1990's.

WSD has been done for general words, for domain specific words, and across different languages. Often, only a few words are disambiguated. For example, Mooney [2] tested 7 different learning algorithms to learn the correct sense of 'line' based on its surrounding words. The order of the words was not taken into account, which is called a bag-of-words approach. Techniques used include naïve Bayes, a perceptron, and decision trees among others. Naïve Bayes was a top performer for both accuracy and required training time. With 1,200 examples, the accuracy was more than 70%, but it was less than 60% with 300 examples. Florian et al. [3] worked with the Senseval2 dataset (www.itri.brighton.ac.uk/events/senseval/) and used an enriched bag-of-words approach that included a weighted bag-of-lemmas and local n-gram context with specific syntactic relations. Their Bayes-based approaches were among the top performers for English (approximately 65% accuracy) and the best for Spanish, Swedish, and Basque. In further studies, they combined classifiers and achieved better accuracy (by about 1%). Pedersen [4] evaluated the use of bigrams for word sense disambiguation. Bigrams are sequences of two words. He tested different methods to select bigrams that occur close to the ambiguous words (within approximately 50 words to the left or right of the ambiguous word) as possible disambiguation features. He tested a decision tree and naïve Bayes classifier. The decision tree with the most accurate disambiguation was based on bigrams selected with a power divergence statistic, which is a goodness-of-fit statistic.

In addition to surrounding words, others drew on information available from external sources such as WordNet (www.cogsci.princeton.edu/~wn/), a general-English lexical reference [5]. Its structure is based on psycholinguistic theories of human memory and includes different senses for words. Inkpen and Hirst [6] used WordNet to disambiguate

[*] This research was performed while the first author was at the National Library of Medicine, Lister Hill National Center for Biomedical Communications

near-synonyms in dictionary entries. They made use of the overlap of words in the dictionary description and WordNet glosses, synsets, antonyms, and polysemy information. They used a decision tree (C4.5) to select the best combination and achieved 83% accuracy.

In biomedicine, word sense disambiguation has been applied to categories of words such as DNA, RNA, and proteins. Hatzivassiloglou et al. [7] used three supervised learning techniques, C4.5 decision trees, naïve Bayes, and inductive learning, and tested different features with an automatically created gold standard to distinguish between genes, proteins, and mRNA. Their best technique, naïve Bayes, achieved 84% accuracy. Liu et al. [8] evaluated different feature sets and classifiers in an extensive study to disambiguate biomedical abbreviations with automatically created gold standards. They trained their classifiers per abbreviation and achieved high accuracy (over 90%) especially when there were thousands of examples from which to learn.

A common element in machine learning techniques is reliance on large datasets. For example, Mooney [2] used 300, 600, and 1,200 examples for training and showed that performance increased with more examples. Some researchers have built gold standards automatically [7, 8] to sidestep this issue. These standards are an excellent approach to comparing different algorithms. However, because they are systematically built, they deviate from the standard human experts would establish. When Hatzivassiloglou [7] asked human experts to assign labels to the same terms as in the artificial gold standard (the disambiguating terms were deleted), the pair-wise agreement of the experts was 78%.

Our goal is to determine whether symbolic knowledge can be used by machine learning algorithms so that they can learn from small, human-created gold standards. The rationale is that by supplying algorithms with additional, external knowledge, comparable to that of experts, fewer examples will be needed for learning. This will be useful since the development of annotated datasets is time-consuming and difficult. We take advantage of the symbolic knowledge in the biomedical domain found in the Unified Medical Language System [9] (UMLS). In addition to using few examples, we also limit the input to what can be found in the sentence containing the ambiguous word. We use the symbolic representation of that sentence in the UMLS Semantic Network [10] and do not use the actual words surrounding the ambiguous term. In this way, our approach, if successful, may augment common bag-of-word approaches.

Methods

Dataset

This study was performed with a dataset provided by the National Library of Medicine (available from http://wsd.nlm.nih.gov/), in which eleven human evaluators disambiguated words occurring in MEDLINE abstracts [11]. The dataset contains 50 English terms, such as cold, mosaic, and growth, which are commonly ambiguous. Each ambiguous term can be mapped to multiple UMLS concepts. For each word, 100 instances were disambiguated by indicating the correct sense with a UMLS concept or the option "None" if no UMLS concept described the correct sense. Each instance is provided with its original MEDLINE abstract. Linguistic and symbolic knowledge is made available for all terms in the entire abstract. MetaMap [12] (available at http://mmtx.nlm.nih.gov/) was used to provide the linguistic information, e.g., part of speech (POS), and to map all terms to UMLS concepts and semantic types. All these mappings are provided in the online dataset.

Our purpose was to train a machine learning technique that can disambiguate the words by choosing the correct mapping. Each mapped concept is also connected to semantic types in the UMLS Semantic Network. We used these semantic types to represent the different meanings of ambiguous terms. For example, based on the UMLS, there are three senses and their related semantic types for "blood pressure." One extra sense is added to be used when none of the previous ones is correct. The UMLS concepts and semantic types are: Blood Pressure (Organism Function), Blood Pressure Determination (Diagnostic Procedure), Arterial Pressure (Laboratory or Test Result), and None of the Above.

Disambiguation Study

We chose a naïve Bayes classifier since it was a top performer in several other word sense disambiguation studies. A naïve Bayes classifier is based on Bayes' probability rules. It takes all presented information into account and is called naïve because it assumes independence between all the features presented to it. We used the Weka software packet to train and test the classifier with 10-fold cross-validation [13].

We report on eight conditions in which symbolic knowledge is cumulatively added to each condition. All knowledge is based on the sentence in which the ambiguous word appears. The intuition is that more complete symbolic information about the ambiguous word, its surroundings or context, and how the word interacts with this context will lead to better disambiguation.

Figure 1 visualizes the relation between the available symbolic knowledge and the experimental conditions. There are two types of basic information about the ambiguous word that we used. The word's status in the phrase: single words or heads of phrases are denoted as main words (MW). We also use the word's part of speech (POS). Four additional types of symbolic knowledge about the ambiguous word's context are evaluated. Phrase types (P-Types) are the semantic types of words in the same phrase as the ambiguous word. Sentence types (S-Types) are the semantic types of all other unambiguous words in the sentence. We believed that additional symbolic knowledge could improve the accuracy of the classifier, and included additional details of the context surrounding the ambiguous word with core (CRel) and non-core (NCRel) relations. These are Semantic Network relations between the unambiguous semantic types found in the sentence. The UMLS Semantic Network has 54 relations that can exist between 135 semantic types. We considered the following seven relations

to be core relations because they closely link concepts in a hierarchical fashion: is a, conceptual part of, consists of, contains, ingredient of, part of, and process of.

Finally, we evaluated how each ambiguous sense fits into its surrounding context. To test this, we added the semantic relations that each ambiguous type can have with its surrounding types (Sense Activation) as a feature to be used by the classifier. The rationale was that the correct sense would have more interaction with the surroundings.

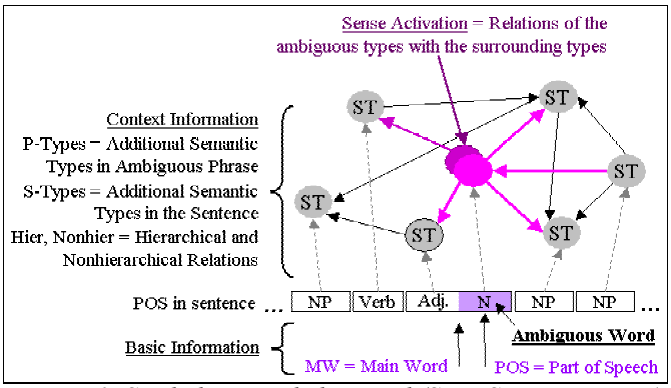

Figure 1: Symbolic Knowledge Used (ST = Semantic Types)

Results

Disambiguation Results

We selected 15 words from the NLM dataset for which the most frequent sense was correct in less than 65% of the instances. This "majority sense performance" served as the baseline for our study. Table 1 provides an overview of the accuracy for each word in each condition and the average improvement. As mentioned above, additional information is added in each condition. For easy reference, we have numbered the conditions, e.g. the baseline is (0). The last two rows in the table provide the results for pair-wise t-test between the experimental conditions and the baseline (Baseline comparison) and between consecutive experimental conditions (Incremental comparison, e.g., 0 vs. 1, 1 vs. 2).

In the first condition (1), we evaluated the importance of the ambiguous word being a single word or head of a phrase (main word) or not. Although the effect was small for most words, for "scale" and "weight," accuracy increased by 14% and 19% compared to the baseline (0). Overall, the improvement in accuracy was significant.

Providing additional information about the ambiguous word's part of speech (2) increased accuracy slightly for two terms (nutrition, repair) but decreased accuracy for several other terms. The differences were not significant.

Adding the semantic types of unambiguous words occurring in the same phrase as the ambiguous word (3) led to increased performance in several cases (man, mosaic, repair, scale, weight, white) but caused deterioration in a few others. Although on average performance improved compared to the previous condition and the baseline, the difference was not significant. When the semantic types of all unambiguous words in the sentence (4) were also available for learning, average accuracy was at its peak (66%). This condition was significantly more accurate than the previous and the baseline. For some words, disambiguation accuracy increased by 20 to 30% compared to the baseline.

To increase the detail of the surrounding context, we added the semantic relations between the unambiguous semantic types that form the context. In conditions 5a the non-core relations are added, while in 6a both core and non-core relation are added. Including information about non-core relations (5a) has a significant adverse affect on accuracy. The core relation information had a small beneficial effect for some words, but the effect was not significant.

Since performance decreased drastically in these conditions, we decided not to pursue them further, but rather to add information about sense activation (5b and 6b) to condition 4. Sense activation consists of the relations the different ambiguous types can have with the unambiguous context. Sense activation based on non-core relations (5b) had a significant adverse effect on accuracy. Core sense activation lowered accuracy for most words compared to condition (4); however, this difference was no longer significant.

Troublesome Instances

Several words responded well to the experimental conditions, while others did not. For example, "repair" had almost 30% increased accuracy in condition 4 compared to the baseline, but the accuracy for "blood pressure" was actually lower in condition 4 than in the baseline.

To find an explanation, we asked whether there was a relation between the baseline performance for each word, the ambiguity in the instances for each word, and the actual accuracy. Figure 2 shows our expectations for accuracy determined by baseline accuracy (part A) or example ambiguity (part B). When the baseline is low, one would expect improvement to be easier to achieve because there are more examples to learn from per sense (the baseline is the maximum percent correct from one sense) and because there is more room for improvement. Similarly, for clear, unambiguous examples, the ambiguity is low and one would expect better learning and so better performance. For troublesome instances, one expects lower performance.

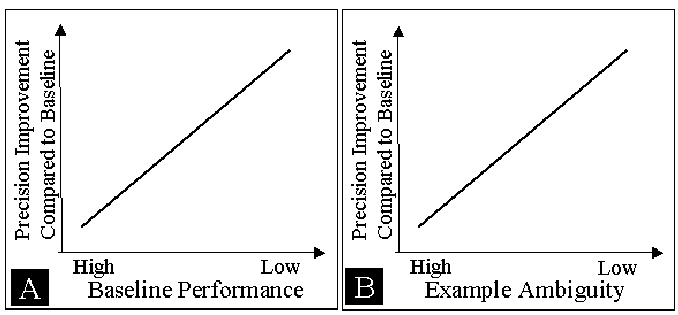

Figure 2: Expected Accuracy Improvement

Table 1: Accuracy of the naïve Bayes classifier for word sense disambiguation

% Accuracy	Baseline	Information Provided to Classifier (Experimental Condition)							
Word	Maj. S. (0)	MW (1)	MW POS (2)	MW POS P-Types (3)	MW POS P-Types S-Type (4)	MW POS P-Types S-Types NC. Rel. (5a)	MW POS P-Types S-Types NC. Rel. H. Rel. (6a)	MW POS P-Types S-Types NC. Sense Act. (5b)	MW POS P-Types S-Types NC. Sense Act C. Sense Act (6b)
Adjustment	62	62	62	62	57	50	51	48	50
Blood pressure	54	54	51	51	46	56	54	48	48
Degree	63	66	66	64	68	60	59	67	70
Evaluation	50	50	50	45	57	53	55	53	54
Growth	63	63	63	62	62	50	50	56	60
Immunosuppression	59	57	57	54	63	61	64	67	65
Man	58	62	58	74	80	62	66	70	70
Mosaic	52	52	46	69	66	42	42	52	56
Nutrition	45	45	53	49	48	37	39	38	40
Radiation	61	61	61	60	72	54	54	63	62
Repair	52	57	58	62	81	68	62	70	69
Scale	65	79	79	82	84	72	71	71	72
Sensitivity	48	54	54	51	70	65	66	70	70
Weight	47	66	66	71	68	54	53	62	59
White	49	49	49	56	62	48	50	59	59
Average	55	58	58	61	66	55	56	60	60
Baseline comparison: t-test, α .05, p-value:		(0 vs.1) .05		(0 vs. 3) < .05	(0 vs. 4) < .005				(0 vs. 6b) < .05
Incremental comparison: t-test, α .05, p-value:		(0 vs. 1) .05			(3 vs. 4) < .05	(4 vs. 5a) < .001		(4 vs. 5b) < .001	

To explore these ideas, we ordered the 15 ambiguous words based on their baseline score (Figure 3) as well as based on the example ambiguity score (Figure 4). We measured the percentage improvement as the improvement in accuracy for the best experimental condition (condition 4, semantic types of unambiguous words in the sentence) compared with the baseline. To calculate example ambiguity we combined the number of senses with a measure of how troublesome each word was to the human evaluators. The NLM dataset contains information about the evaluation of all 100 instances of each word by the eleven experts. In some cases, the experts did not agree on the correct sense of a word and only chose one sense after extensive discussion. Those requiring discussion are reported as "unresolved counts." We labeled words with many senses and troublesome examples (numbers were multiplied) as words with high example ambiguity.

Figure 3 shows that the actual performance improvement for the words ordered by their baseline performance. There is no improvement with a lower baseline (no significant correlation). However, actual performance seems to increase when the example ambiguity is lower (Figure 4). Although this is a small test set, a trend can be seen for words with higher example ambiguity (left side) to have lower performance scores and words with lower example ambiguity to have higher performance scores. We tested the correlation with the Pearson coefficient and found a strong trend (one-tailed, r = -0.379, p = 0.8). If we exclude the first word (mosaic), the correlation is significant (one-tailed, r = -0.725, p < .01).

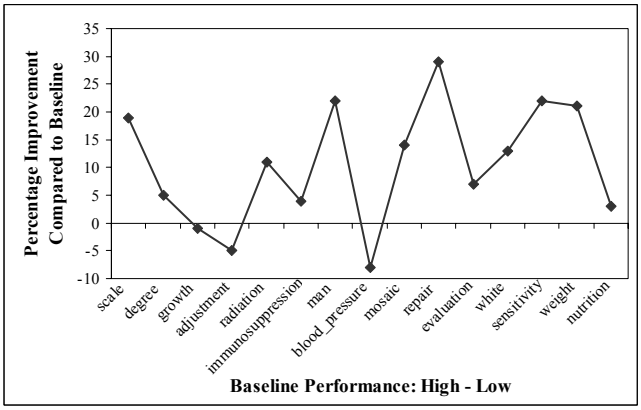

Figure 3: Actual Accuracy Improvement (Baseline-Ordered))

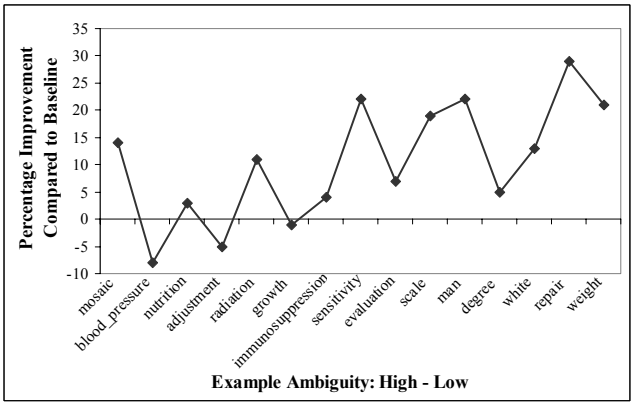

Figure 4: Actual Accuracy Improvement (Ambiguity-Ordered)

Discussion

We assumed that more symbolic information would be better, but this was not the case. The class of non-core relations had a negative effect when they were included in the context information. We plan a more detailed evaluation for individual semantic relations to look into this effect. We will also look at the interaction of the symbolic knowledge with different machine learning approaches. A few samples showed that the results are different with decision trees.

Although no information about the human evaluators agreement was provided to naïve Bayes classifier, there was a trend that better accuracy was achieved with less troublesome instances. This indicates that gold standards developed by multiple experts display variability and inconsistencies. It would be interesting if the classifier could learn to classify for each individual experts.

Conclusions

The purpose of this study was to discover if symbolic knowledge can be used by machine learning algorithms so that it can be added to the common, example-based approach and allow learning on smaller datasets. We used a naïve Bayes classifier to disambiguate medical terms and the UMLS for its symbolic knowledge. Only information from the sentence in which the ambiguous word appeared was used.

We tested 8 different experimental conditions and compared them with the majority sense baseline. In each condition more information was provided to the naïve Bayes classifier. However, it was not the condition with the most information that resulted in the best performance. Three types of information helped accuracy: information about the word being the main word or not, UMLS semantic types associated with unambiguous words in the sentence, and core relation between the context and the ambiguous senses. When evaluating the potential causes for the high variability between the performances of different words, we discovered an unexpected trend related to example ambiguity. Words that were troublesome to the human evaluators were generally also harder to automatically disambiguate.

We conclude that using symbolic knowledge for word sense disambiguation is a promising approach. Future work will include combining and testing other machine learning techniques and comparing the common approach (using the surrounding words) with and without the symbolic knowledge.

Acknowledgements

The authors thank the experts from the National Library of Medicine who evaluated the medical terms. We also thank Halil Kilicoglu and Jim Mork for their assistance in making the dataset accessible, and Olivier Bodenreider for his suggestions. The first author is grateful to Alexa McCray for the invitation to the Lister Hill National Center for Biomedical Communications.

References

[1] Ide N, Véronis J. Word sense disambiguation: The state of the art. Computational Linguistics 1998;24(1):1-41.

[2] Mooney RJ. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In: Conference on Empirical Methods in Natural Language Processing; 1996; 1996.

[3] Florian R, Cucerzan S, Schafer C, Yarowsky D. Combining classifiers for word sense disambiguation. Natural Language Engineering 2002;1(1):1-14.

[4] Pedersen T. A decision tree of bigrams is an accurate predictor of word senses. In: Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics; 2001; 2001. p. 79-86.

[5] Miller GA, Beckwidth R, Fellbaum C, Gross D, Miller K. Introduction to wordnet: An on-line lexical database. 1998.

[6] Inkpen DZ, Hirst G. Automatic sense disambiguation of the near-synonyms in a dictionary entry. In: 4th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2003); 2003 February; Mexico City; 2003. p. 258-67.

[7] Hatzivassiloglou V, Duboué PA. Disambiguating proteins, genes, and rna in text: A machine learning approach. Bioinformatics 2001;1(1):1-10.

[8] Liu H, Lussier YA, Friedman C. Disambiguating ambiguous biomedical terms in biomedical narrative text: An unsupervised method. Journal of Biomedical Informatics 2001;34:249-61.

[9] Humphreys B, Lindberg D, Schoolman H, Barnett G. The unified medical language system: An informatics research collaboration. Journal of the American Medical Informatics Association 1998;5(1):1-11.

[10] McCray A. Representing biomedical knowledge in the umls semantic network. In: NC B, editor. High-performance medical libraries: Advances in information management for the virtual era. Westport, CT: Meckler Publishing; 1993. p. 45-55.

[11] Weeber M, Mork J, Aronson A. Developing a test collection for biomedical word sense disambiguation. In: AMIA Symposium; 2001; 2001. p. 746-50.

[12] Aronson A. Effective mapping of biomedical text to the umls metathesaurus: The metamap program. In: AMIA Symp.; 2001; 2001. p. 17-21.

[13]Witten IH, Frank E. Data mining: Practical machine learning tools and techniques with java. San Francisco: Morgan Kaufmann; 2000.

Address for correspondence

Gondy Leroy, School of Information Science, Claremont Graduate University, 130 E. Ninth Street, Claremont CA 91711, gondy.leroy@cgu.edu